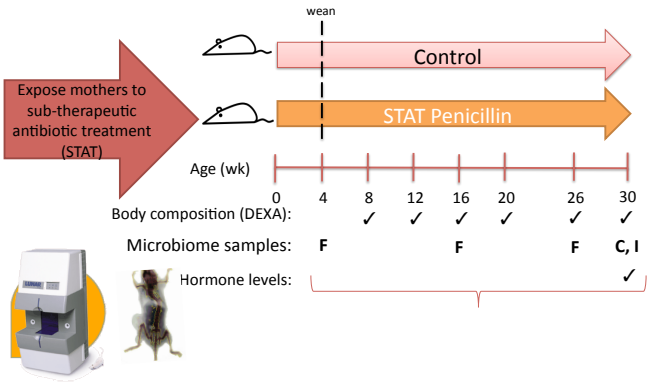
Testing and multiple testing

Study design



Questions

- Are any taxa associated with antibiotic treatment?
- Is there difference in microbiome composition over time (within/between treatments)?
- Is there correlation between abundance of any taxa and metabolic phenotypes?
- Are there any pairwise correlations between taxa?

Hypotheses

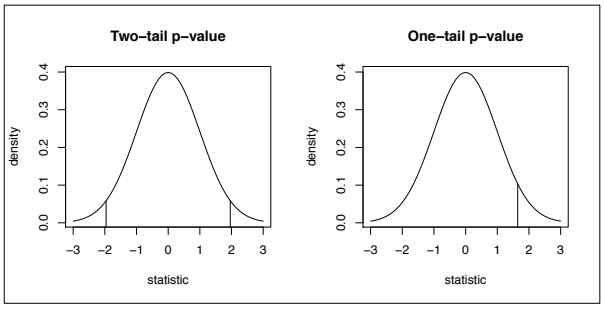
- Are precise statements that are amenable to being proven false using data.
- Null hypothesis: a proposition that corresponds to default position. ("Nothing special is happening")
- Alternative hypothesis: a proposition that describes a non default outcome ("Something is going on")
- The inference is obtained by rejecting the Null hypothesis. Null hypothesis can never be confirmed by the data!

Example of hypotheses

- General question: Are any taxa associated with antibiotic treatment?
- Univariate hypothesis question: Is taxon T associated with antibiotic treatment?
- Null hypothesis: abundance of taxon T follow the same distribution in treated and control groups.
- Alternative hypothesis 1: abundance of taxon T follow distribution of different *form* in the two groups.
- Alternative hypothesis 2: abundance of taxon T follow the same form of distribution but with different *mean* between groups.
- Alternative hypothesis 3: abundance of taxon T follow the same form of distribution but with different *median* between groups.
- Alternative hypothesis 4: abundance of taxon T follow the same form of distribution but with different *variance* between groups.

P-values

- If the Null Hypothesis was in fact true a *statistic* used to perform the test would follow a certain distribution: null distribution.
- P-value is the tail probability under the null distribution.



Distribution of OTU abundance data

- *Justifiable* distribution assumptions often allow for better statistical tests
- Properties of OTU abundance data:
  - Correlated: Sums to 1, hence to increase something, something else has to decrease
  - Variable across subjects
- Can *possibly* be modeled through compound Dirichlet-Multinomial distribution; however, useful marginal univariate tests have not yet been derived:
  - http://www.amstat.org/meetings/jsm/2011/onlineprogram/AbstractDetails.cfm?abstractid=302334
- Have to rely on non-parametric (distribution free) tests, possibly at the cost of decreasing the power of the tests

Mann-Whitney U or Wilcoxon rank-sum two-sample test

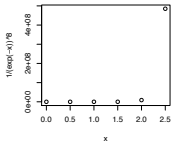
- Assumptions:
  - Independent observations
  - Observations can be ordered with respect to each other
- Null hypothesis: The distribution in two samples is the same. If one randomly draws one observation from each sample X, Y; then Pr(X>Y) = Pr(Y>X)
- Two-sided alternative hypothesis: Pr(X>Y) ≠ Pr(Y>X)
- Interpretation: for continuous observations, significant tests indicate change in the median
- Example: Is the abundance of a taxon different between STAT and Control?
- In R: `wilcox.test`

Kruskal-Wallis one-way analysis of variance (more than two samples/groups)

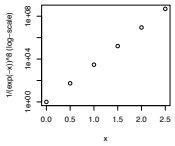
- Assumptions:
  - Independent observations that follow distribution with the same shape and scale
  - Observations can be ordered with respect to each other
- Null hypothesis: The location (median) of all the groups is the same.
- Alternative hypothesis: Location for at least one group is different from location of at least one other group
- Example: Is the aundance of a taxon different in STAT/ control over 3 sampled time points?
- In R: `kruskal.test`

Correlation coefficients, rank correlations

- Linear correlation coefficient (Pearson) assumes linear dependence between two variables
- Rank correlation coefficient measure the extent of monotonicity between two variables
- Null hypothesis for correlation testing: correlation coefficient is equal to 0.



Pearson correlation coefficient: 0.66 (not significant, p=0.15)



Diaconis, P. (1988), Group Representations in Probability and Statistics, Lecture Notes-Monograph Series, Hayward, CA: Institute of Mathematical Statistics, ISBN 0-940600-14-5

Rank correlation coefficients

- Spearman's ρ: Rank correlation measure defined as the Pearson correlation of the two variables after conversion to ranks
- Kendall's τ: Rank correlation measure based on counting concordant pairs. [(x1,y1) and (x2,y2) are concordant if x1>x2 when y1>y2]
- Example: Is there correlation between any given two taxa? Is there correlation between a given metabolic variable and a given taxon?
- In R:
  - `cor.test(x, y, method='spearman')`
  - `cor.test(x, y, method='kendall')`

Problems with testing many hypotheses simultaneously

- We have many OTUs that we would like to apply the test to.
- If the test is applied at specified significance level (probability of falsely rejecting the null, when it is true), we cannot guarantee that combined result is at the significance level originally specified.
- Expected number of rejections by mere chance $m*\alpha$
- How do we control significance for multiple tests?

FWER: Familywise error rate

	# not-rejected	# rejected	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m-m_0$
Total	m-R	R	m

FWER control methods adjust (more stringent tests need to be performed) the significance of each individual test to ensure overall significance at given α.

- Suppose we perform m tests (e.g. m taxa are tested for association with antibiotic treatment)
- The number of true null hypotheses is unknown m_0
- V is false positive rate (Type I error)
- T is false negative rate (Type II error)
- We observe R, but S, T, U, V are unobserved
- FWER = Pr(V≥1)

Example: Bonferroni correction

- To ensure overall significance at a given α, one performs each individual test at α' = α/m
- Very stringent, results in loss of power (increase in Type II error)

FDR: false discovery rate

- Modifies the idea of controlling Type I error, to instead control the rate at which type I errors do occur
- FDR is the expected value of V/R

Methods for FDR control

- Benjamini–Hochberg
  - Assumes tests are independent
- Benjamini–Hochberg–Yekutieli
  - Assumes that tests are uniformly correlated:
    - Positively correlated: if one test has low p-value, other tests are *more* likely to also be significant
    - Negatively correlated: if one test has low p-value, other tests are *less* likely to be significant

FDR in R

- FDR is implemented in R as a p-value adjustment procedure.
- Input: p-values for a set of univariate tests
- Output: p-values that are adjusted to FDR
- E.g. 0.05 adjusted p-value means that expected rate of false positives is 0.05 for tests significant at that adjusted level
- `p.adjust`
  - Methods:
    - `method = 'fdr':` Benjamini-Hochberg
    - `method = 'BY':` Benjamini-Hochberg-Yekutieli